Projecto GREASE-II

Actividades em curso no IST

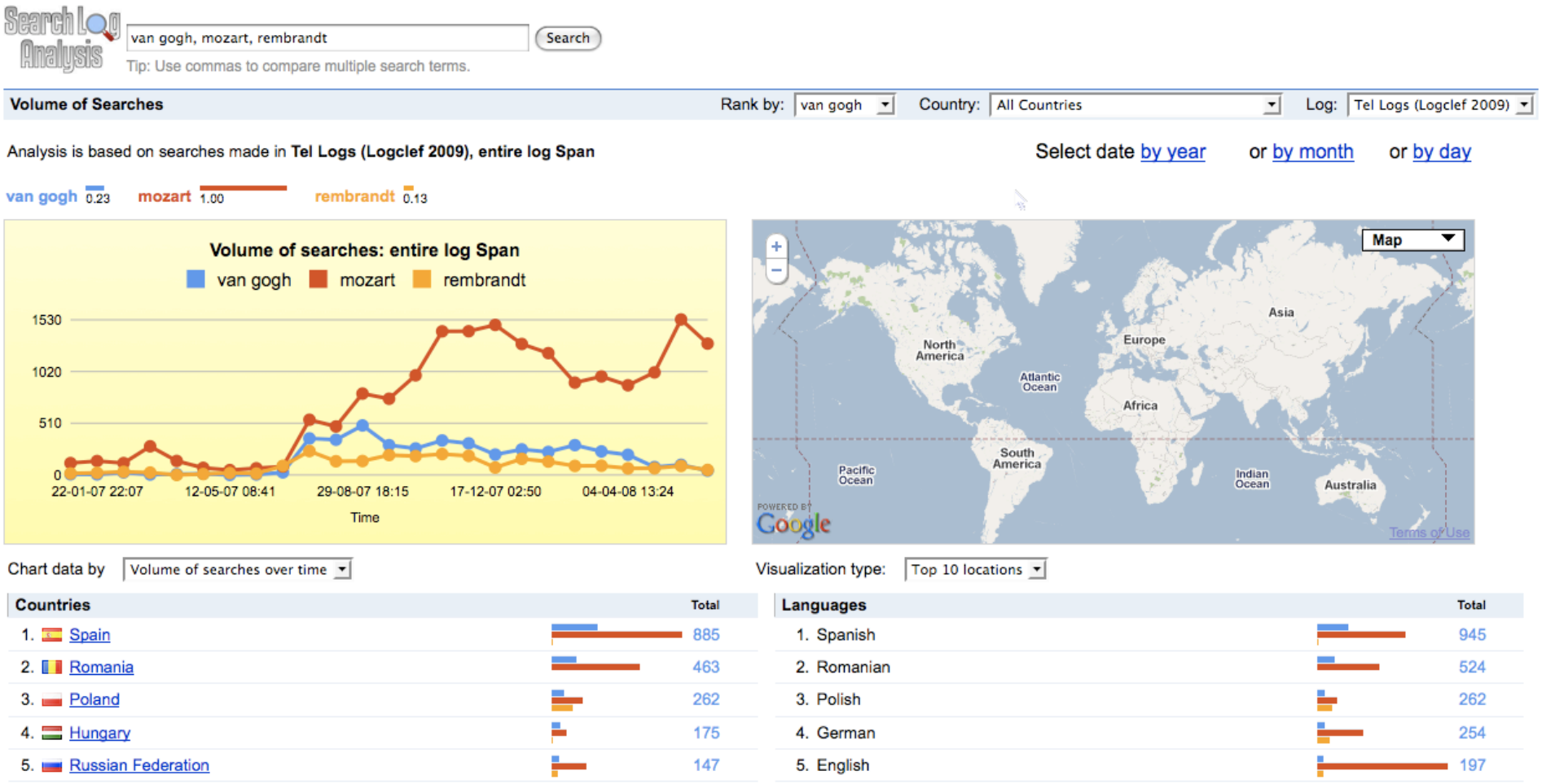
Investigação recente

- **Classifying pages according to locational relevance**
  - Anastácio, Martins e Calado - Artigo aceite no EPIA 2009
  - Support Vector Machines para classificar artigos ODP como "locais" ou "globais"
- **Comparing approaches for assigning documents to geo-scopes**
  - Anastácio, Martins e Calado – Artigo aceite no INFORUM 2009
  - Comparação de vários métodos, incluíndo Web-a-Where e PageRank
  - Provavelmente será extendido para artigo de revista
- **Using geographic scopes for contextual advertisement**
  - Anastácio, Martins e Calado – Artigo submetido ao CIKM 2009
  - Foco da tese de mestrado do Ivo Anastácio
  - Retrieval de anúncios com base nos âmbitos geográficos e na "locational relevance" das páginas
- **A machine learning approach for gazetteer record linkage**
  - Martins – Artigo submetido ao ACMGIS 2009
  - SVMs e Decision Trees para classificar pares de registos de um gazetteer como dup. ou non-dup.
  - Provavelmente será extendido para artigo de revista
- **Learning to rank for geographic information retrieval**
  - Martins e Calado – Em progresso (a submeter ao WSDM 2009)
  - Usar SVM-map sobre dados (colecções em Inglês) das 4 edições do GeoCLEF

Outras actividades em curso

- Vários trabalhos de mestrado, em diferentes estágios:
  - **Spatio-Temporal Search Log Analysis**
    - Aluno de mestrado a terminar, co-orientação com Pável Calado
  - **Reconhecimento e desambiguação de entidades geo. e temporais**
    - Dois alunos de mestrado a começar, co-orientação com Luísa Coheur
  - **User Interfaces for Geographic Information Retrieval**
    - Aluno de mestrado a começar, co-orientação com Manuel Fonseca
    - Aluno de mestrado focado em apresentações sobre mapas, co-orientação com Borbinha
    - Eu ando a explorar clustering de resultados
  - **Hierarchical classification of large collections with few training data**
    - Aluno de mestrado a começar, co-orientação com o Pável Calado e Daniel Gomes
  - **Geographic recommender systems**
    - Aluno de mestrado a começar, co-orientação com o Andreas Wichert
    - Explorar Tied Boltzman Machines, TASTE e informação na Web

Alguns resultados
(spatio-temporal search log analysis)



Alguns resultados
(gazetteer record linkage)

	Support Vector Machines					Alternating Decision Trees				
	Precision	Recall	F_1	Accuracy	Error	Precision	Recall	F_1	Accuracy	Error
Place names	0.993	0.954	0.973	97.3787	2.6213	0.988	0.971	0.979	97.9496	2.0504
Footprints	0.797	0.941	0.863	85.0506	14.9494	0.944	0.95	0.947	94.6535	5.3465
Names + footprints	0.992	0.958	0.975	97.5603	2.4397	0.989	0.976	0.983	98.2611	1.7389
All	0.992	0.962	0.977	97.6901	2.3099	0.987	0.979	0.983	98.313	1.687

Table 2: Performance comparison between classification algorithms and different feature vectors.

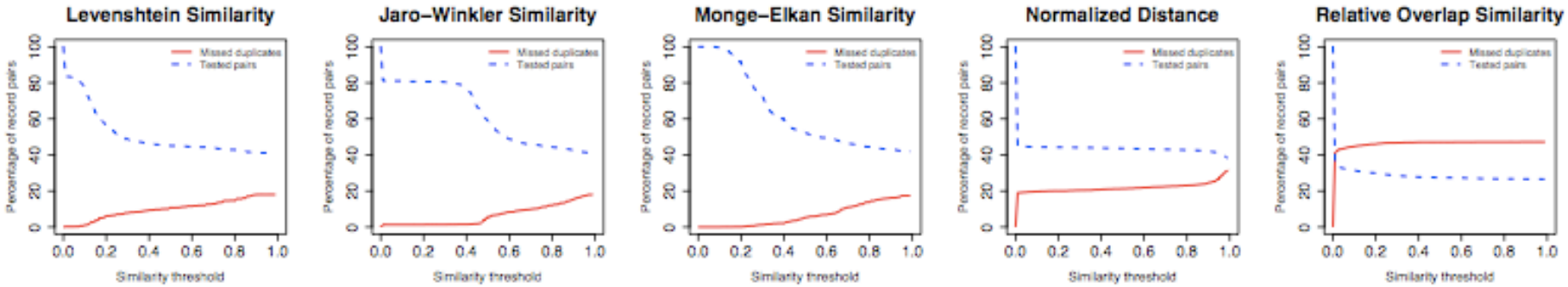


Figure 2: Effect of pre-filtering pairs with basis on simple similarity scores.

Alguns resultados
(metodos de atríbuição de scopes)

	Avg. Distance	Std.dev. Distance	Avg. Overlap	Std.dev. Overlap	Accuracy (D=0 Km)	Accuracy (D<100 Km)	Accuracy (O>0.75)
Placemaker Admin.	1031	**1460**	0.42	0.49	0.38	0.45	0.38
Web-a-Where	**953**	1880	**0.51**	0.49	**0.48**	0.56	**0.48**
GIPSY	1267	2249	0.25	0.41	0.21	0.44	0.23
Covering Area	2656	3009	0.25	**0.38**	0	0.21	0.18
Most Frequent	1093	2331	0.49	0.49	**0.48**	**0.57**	**0.48**
Non-outliers	1740	2826	0.36	0.46	0.25	0.39	0.34

Table 2. Comparison of human-assigned versus automatically assigned scopes.

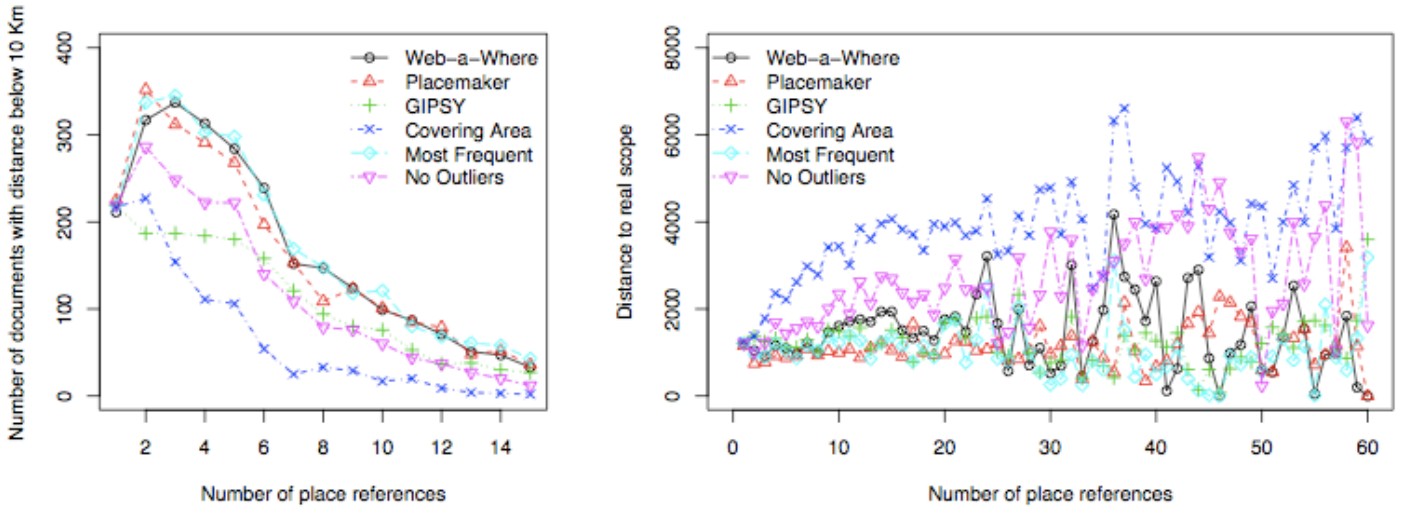


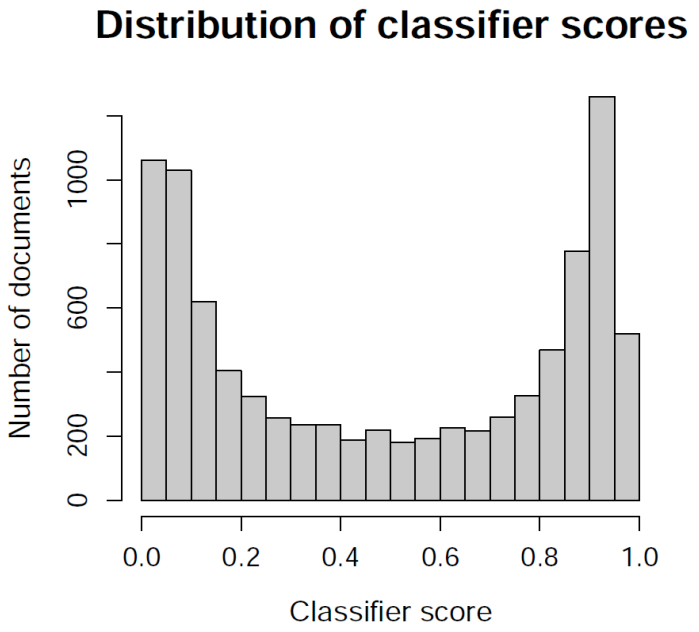
Fig. 2. Correlation between number of place references and scope accuracy.

Alguns resultados (Relevância geográfica de documentos)

- Dados de teste/treino : 8000 páginas do ODP com referências geográficas
  - Classe GLOBAL
    - 2000 fora da categoria Regional
    - 2000 abaixo de USA
  - Classe LOCAL
    - 2000 abaixo de [State]
    - 2000 abaixo de [City]

- SVM Classifier, 10-fold cross-validation

- Features usadas no classificador:
  - Texto (Accuracy = 80%)
  - Geo (Accuracy = 81%)
  - Texto+Geo (Accuracy = **89%**)



Caminhos a explorar

- **Continuar investigação em abordagens learning to rank:**
  - Explorar utilização do SVM-comb (ou SVM-ndcg)
  - Explorar semi-supervised machine learning approaches
  - Explorar feature selection

- **Usar machine learning para resolver entidades geográficas e temporais em documentos:**
  - Usar SVMs (ou HMMs ou ainda CRFs) para reconhecimento.
  - Usar learning to rank para desambiguar entidades, combinando várias formas de evidência (default senses + spatial and semantic minimality).

- **Extracção de informação e *spatio-temporal topic modeling***
  - Intersecção entre GIR, extracção de informação e Topic Detec. and Tracking
  - Provavelmente o doutoramento do Ivo irá evoluir neste sentido

Ideias para colaboração com a FCUL

- **Spatio-temporal search log analysis and topic modeling**
  - Intersecção com outros projectos vossos.
  - Contactos com investigadores do grupo do Timos Sellis (cozinhar projecto novo)

- **Learning to rank**
  - Miguel Costa anda a explorar isto para tarefas de temporal IR

- **Reconhecimento e desambiguação de entidades**
  - Explorar machine learning ou colecções de n-gramas (ideia: usar colecção da FCCN)
  - Precisamos de mais dados, nomeadamente da colecção SpatialML.
  - Contactos com investigadores dos grupos da Nieves Brisaboa e do Timos Sellis

- **Integração de dados no contexto de gazetteers**
  - Francisco Javier anda a trabalhar com a GeoNET-PT
  - Contactos com o IGP

- **Spatio-temporal XML processing**
  - Tenho olhado para isto no contexto do meu ensino no IST
  - Aluno de mestrado a começar com o Pável Calado
  - Contactos com investigadores do LATINGeo (Madrid)

Conhecimentos técnicos adquiridos

Temos *know-how* relevante na utilização de várias ferramentas essenciais para realizar experiências na área:

- Machine Learning (utilização do software **WEKA**)
- Text Mining (utilização do **LingPipe**)
- GeoParsing (serviços do **SAPO, Yahoo! PlaceMaker e Metacarta**)
- Gazetteers (utilização do **geonames e do ADL**)
- Processamento de XML (utilização de **motores XQuery**)
- Learning to Rank (utilização do **SVM-map**)
- Google Maps
- **MySQL e PostgreSQL** full-text and spatial capabilities
- Medição de similaridade (utilização do software **SIMPACK**)